KAMPUS AKADEMIK PUBLISING
Jurnal Sains Student Research (JSSR)
Vol.4, No.6 Desember 2026
e-ISSN: 3025-9851, p-ISSN : 3025-986X, Hal 1156-1171
DOI : https://doi.org/10.61722/jssr.v4i6.13255

Imputasi Data Menggunakan Teknik Statistik Hot-Deck Untuk Klasifikasi Penyakit Menular HIV Dengan Metode Random Forest

Anida Fitri Hardiantari[1*], Fajri Rakhmat Umbara[2], Ridwan Ilyas[3]
[1]Program Studi Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
[2]Program Studi Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
[3]Program Studi Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
*Penulis Korespondensi: anidafitri22@if.unjani.ac.id

Abstract. *Infectious diseases caused by microorganisms, such as the Human Immunodeficiency Virus (HIV), remain a significant global health issue. In health data management, the problem of missing values frequently arises, potentially compromising data quality and affecting analysis results and classification processes. Therefore, appropriate methods for handling missing values are essential to maintain data quality. This study aims to implement the Hot-deck Imputation method to address missing values in an HIV dataset and to evaluate classification performance using the Random Forest algorithm. The dataset was obtained from Harvard Dataverse and split into two subsets: 80% for training and 20% for testing. Model performance was evaluated using several metrics: Accuracy, Precision, Recall, and F1-Score. The evaluation results demonstrate that the Hot-deck Imputation method successfully handled missing values, enabling the dataset to be used for classification with the Random Forest algorithm. Testing results showed an Accuracy of 56.07%, Precision of 54.86%, Recall of 63.58%, and an F1-Score of 58.90%. These results serve as a pre-processing step for handling missing values prior to the classification process for the HIV dataset.*

Keywords: *Missing Value; Hot-deck Imputation; Random Forest; Classification; HIV.*

Abstrak. Penyakit menular yang disebabkan oleh mikroorganisme, seperti *Human Immunodefiency Virus* (HIV), masih menjadi salah satu permasalahan kesehatan yang signifikan di dunia. Dalam pengelolaan data kesehatan, sering ditemukan permasalahan *missing value* yang dapat menurunkan kualitas data serta memengaruhi hasil analisis dan proses klasifikasi. Oleh karena itu, diperlukan metode penanganan *missing value* yang tepat agar kualitas data tetap terjaga. Penelitian ini bertujuan untuk mengimplementasikan metode Hot-deck Imputation dalam menangani *missing value* pada dataset HIV serta mengevaluasi performa klasifikasi menggunakan algoritma Random Forest. Dataset yang digunakan berasal dari Harvard Dataverse kemudian dipisahkan menjadi dua bagian yaitu, data latih sebesar 80% dan data uji 20%. Selanjutnya, kinerja model dievaluasi berdasarkan beberapa metrik, meliputi Accuracy, Precision, Recall dan F1-Score. Berdasarkan hasil evaluasi yang diperoleh, penelitian ini menunjukkan bahwa metode Hot-deck Imputation berhasil menangani *missing value* pada dataset sehingga data dapat digunakan dalam proses klasifikasi menggunakan algoritma Random Forest. Berdasarkan hasil pengujian, model memperoleh nilai Accuracy sebesar 56,07%, Precision sebesar 54,86%, Recall sebesar 63,58%, F1-Score sebesar 58,90%. Hasil digunakan sebagai tahap *pre-processing* dalam menangani *missing value* sebelum proses klasifikasi pada dataset HIV.

Kata kunci: Missing value; Hot-Deck Imputation; Random Forest; klasifikasi; HIV.

1. LATAR BELAKANG

Penyakit infeksi yang disebabkan oleh *mikroorganisme* masih menjadi salah satu masalah kesehatan yang banyak ditemukan di berbagai negara. Salah satu penyakit

Naskah Masuk:15 Oktober 2026; Revisi: 16 Oktober 2026; Diterima: 17 Oktober 2026; Terbit: 31 Oktober 2026

menular tersebut adalah infeksi yang disebabkan oleh Human Immunodeficiency Virus (HIV). HIV merupakan virus yang menyerang sistem kekebalan tubuh, terutama sel darah putih (limfosit), sehingga kemampuan tubuh dalam melawan infeksi menurun. Kondisi ini menyebabkan penderitanya lebih rentan mengalami berbagai penyakit, sulit pulih dari infeksi oportunistik, dan pada kondisi tertentu dapat berujung pada kematian. Infeksi HIV dapat terjadi pada berbagai kelompok usia, baik anak muda maupun orang dewasa. Penularan HIV berlangsung melalui paparan cairan tubuh yang mengandung virus, seperti darah, air mani, cairan vagina, air susu ibu, hubungan seksual, hubungan sesama jenis, penggunaan jarum suntik yang terkontaminasi, transfusi darah atau produk darah yang terinfeksi, serta penularan dari ibu yang hidup dengan HIV kepada bayinya (Baratawidjaja, 2024).

Dalam beberapa tahun terakhir, pengelolaan data kesehatan, khususnya pada penyakit menular seperti HIV, semakin berkembang seiring meningkatnya pemanfaatan data dalam bidang kesehatan. Namun, data yang dikumpulkan sering kali memiliki permasalahan berupa missing value. Kondisi ini dapat disebabkan oleh berbagai faktor, seperti responden yang tidak dapat ditemui, data yang tidak tercatat oleh petugas, maupun kesalahan pada saat proses pengumpulan data (Little & Rubin, 2019). Keberadaan missing value dapat menyebabkan informasi pada dataset menjadi tidak lengkap sehingga perlu ditangani sebelum dilakukan proses analisis atau klasifikasi (Han, Kamber, & Pei, 2016)

Imputasi merupakan proses untuk menentukan dan menetapkan nilai pengganti terhadap missing value sehingga dataset menjadi lebih lengkap dan dapat digunakan dalam proses analisis (Tan, Steinbach, & V.Kumar, 2016). Salah satu metode imputasi yang digunakan dalam penelitian ini adalah Hot-deck Imputation. Metode Hot-deck dipilih karena dapat diterapkan pada berbagai jenis data serta digunakan untuk menangani missing value tanpa mengurangi jumlah data yang tersedia (Fadillah & S.Muchlisoh, 2002).

Penelitian Rhaudatunnisa dan Wilantika (2021) melakukan pembentukan nilai hilang pada lima dataset, yaitu SUSENAS Kor Maret 2019, Iris, E.Coli, Breast Cancer 1, dan Breast Cancer 2, dengan persentase nilai hilang mulai dari 5% hingga 50%. Pembentukan nilai hilang tersebut dilakukan dengan memperhatikan mekanisme *missing value,* yaitu MCAR, MAR, dan MNAR (Rhaudatunnisa & Wilantika, 2019).

Dataset yang telah memiliki nilai hilang kemudian diimputasi menggunakan tiga metode, yaitu Hot-deck Imputation, K-Nearest Neighbor Imputation (KNNI), dan Predictive Mean Matching (PMM). Kinerja ketiga metode tersebut dievaluasi menggunakan nilai RMSE, UCE, SCE, serta waktu komputasi yang digunakan (Rhaudatunnisa & Wilantika, 2019). Dari hasil evaluasi, perbandingan antara KNNI dan Hot-deck Imputation menunjukkan adanya *trade-off* yang cukup jelas antara akurasi dan efisiensi waktu komputasi. Dari sisi akurasi, KNNI secara konsisten menghasilkan nilai RMSE yang lebih rendah dibandingkan Hot-deck Imputation pada beberapa dataset, dengan nilai RMSE KNNI berada pada kisaran 14, sedangkan Hot-deck berada pada kisaran 32.

Namun demikian, berdasarkan analisis skoring secara keseluruhan terhadap ketiga metode, Hot-deck Imputation memperoleh skor tertinggi dan ditetapkan sebagai metode terbaik dalam menangani *missing value,* terutama karena keunggulannya dalam efisiensi waktu dan komputasi dan konsistensinya pada dataset dengan jumlah variabel yang lebih banyak (Rhaudatunnisa & Wilantika, 2019).

Pada penelitian sebelumnya, evaluasi metode imputasi lebih berfokus pada kesalahan estimasi dan waktu komputasi, namun belum mengkaji penerapan metode Hot-deck Imputation sebagai tahap pre-processing sebelum proses klasifikasi menggunakan algoritma Random Forest pada dataset penyakit menular. Oleh karena itu, penelitian ini dilakukan untuk mengimplementasikan metode Hot-deck Imputation dalam menangani missing value serta mengevaluasi performa model Random Forest pada dataset HIV.

2. KAJIAN TEORITIS

A. *Hot-Deck Imputation*

Hot-deck Imputation merupakan salah satu metode imputasi yang digunakan untuk menangani *missing value* dengan mengganti nilai yang hilang menggunakan nilai dari observasi lain yang memiliki data lengkap pada dataset yang sama. Observasi yang menyediakan nilai pengganti disebut sebagai *donor,* sedangkan observasi yang memiliki nilai yang hilang disebut sebagai *recipient.* Metode ini bertujuan untuk mempertahankan jumlah data sehingga proses analisis tetap dilakukan tanpa menghapus data yang memiliki *missing value* (van Buuren, 2018).

Pada metode Hot-deck Imputation, nilai pengganti diperoleh dari data donor yang

berasal dari dataset yang sama. Pemilihan donor dapat dilakukan dengan berbagai pendekatan, seperti berdasarkan kelompok tertentu, karakteristik yang serupa, maupun secara acak (random donor), bergantung pada rancangan metode yang digunakan. Setelah donor dipilih, nilai dari donor tersebut digunakan untuk menggantikan nilai yang hilang pada data penerima (recipient) sehingga data menjadi lebih lengkap untuk proses analisis selanjutnya (Little & Rubin, 2019).

B. *Random Forest*

Random Forest merupakan suatu algoritma dalam melakukan klasifikasi yang termasuk *ensambel learning* (Dedot, 2023). Pada algoritma ini terdapat pohon *k* pohon dengan vektor random yang independen dengan vektor-vektor random sebelumnya, tetapi memiliki distribusi yang identik, juga menyatakan bahwa metode ini memanfaatkan algoritma *decision tree* dalam melakukan klasifikasi. Kemudian dibentuk sebuah model dengan menerapkan metode *bootsrap aggregating* ketika membentuk sebuah sampel *training set,* dan setiap *tree* yang dibentuk menggunakan metode yang sama untuk membangun CART (*Classification and Regression Tree)* (Efron, Halloran, & Holmes, 2026).

Dalam penggunaan metode *random forest,* diperlukan variabel prediktor yang digunakan sebagai prediksi dalam pembentukan *decision tree* pada model *random forest.* Pentingnya suatu variabel prediktor dilihat dari nilai skor *importansi* (Efron et al., 2026). Semakin tinggi skornya, maka semakin penting variabel prediktor tersebut. Kemudian untuk variabel prediktor yang digunakan merupakan variabel yang berhubungan dengan variabel yang akan diprediksi (Delen, 2026).

C. *Human Immunodeficiency Virus*

Human Immunodefienciency Virus merupakan salah satu penyakit yang disebabkan virus penyerang imunitas yang dapat ditularkan dari suatu penderita ke orang lain, yang hingga saat ini belum ditemukan obat untuk menyembuhkannya secara total (Centers for Disease Control and Prevention, 2026). Masa inkubasi HIV 6 bulan sampai 5 tahun, *Window period* selama 6-8 minggu adalah keadaan dimana tubuh telah terinfeksi HIV dan belum bisa didekteksi oleh pemeriksaan loboratorium, seorang dengan HIV dapat bertahan sampai dengan 5 tahun, jika tidak diobati maka penyakit ini akan bermanifestasi sebagai AIDS. Gejala klinis muncul sebagai penyakit yang tidak khas seperti : Diare, Kandiasis mulut yang luas, *Pneumonia interstisialis limfositik, Ensefalopati kronik*

(Bertha, 2019).

3. METODE PENELITIAN

Metode penelitian yang diterapkan dalam penelitian ini mencakup tahapan pengumpulan data, pre-processing data, pembagian data menjadi data latih dan data uji, imputasi missing value menggunakan metode Hot-Deck, penerapan algoritma Random Forest untuk proses klasifikasi, serta evaluasi kinerja model. Data yang digunakan pada penelitian ini berasal dari Harvard Dataverse dengan judul "Behavioural and drug-taking risk behaviour among female sex workers and men in mobile occupations in Indonesia, 2002-2004". Dataset ini memuat data survei terkait perilaku risiko dari dua kelompok populasi utama, yakni pekerja seks komersial perempuan dan pria dengan mobilitas tinggi di Indonesia pada rentang waktu 2002 hingga 2004. Urutan tahapan pre-processing dalam penelitian ini adalah Seleksi atribut & label, Pemisahan atribut & label, Cek & hapus label kosong, dan Data Cleaning (Penggunaan Kode Tidak Valid). Pembagian data menggunakan metode stratified sampling. Metode yang digunakan pada penelitian ini adalah Hot-deck Imputation dengan pendekatan random donor. Proses klasifikasi dilakukan menggunakan algoritma Random Forest setelah seluruh tahapan pre-processing dan imputasi selesai. Setelah proses klasifikasi, dilakukan evaluasi performa model menggunakan Confusion Matrix dan beberapa metrik evaluasi.

4. HASIL DAN PEMBAHASAN

A. Hasil Pre-Processing

Seleksi Fitur

Tahap seleksi fitur dilakukan untuk menentukan atribut yang relevan dan memiliki keterkaitan dengan perilaku serta faktor risiko penularan HIV dari keseluruhan variabel yang tersedia pada dataset survei. Proses seleksi fitur dalam penelitian ini mengacu pada penelitian terdahulu yang menggunakan dataset yang sama, yaitu penelitian oleh Ardiansyah (2025). Penelitian tersebut mengidentifikasi beberapa variabel yang memiliki hubungan dengan perilaku berisiko terkait penularan HIV, di antaranya konsistensi penggunaan kondom, riwayat infeksi menular seksual, persepsi terhadap risiko HIV, serta karakteristik demografis responden. Oleh karena itu, atribut yang digunakan dalam penelitian ini dipilih berdasarkan kesesuaiannya dengan tujuan penelitian serta didukung oleh temuan pada penelitian terdahulu.

Berdasarkan hasil kajian literatur dan penelitian terdahulu, serta mempertimbangkan keterkaitannya dengan variabel target (*testresults*), dipilih tujuh atribut prediktor yang dinilai relevan untuk digunakan dalam proses klasifikasi, yaitu *sex*, *age*, *educ*, *consiscondomcs*, *anystd*, *feelriskcondom*, dan *anyintervention*. Pemilihan atribut tersebut didasarkan pada hubungan konseptual maupun empiris dengan faktor-faktor yang berkaitan dengan risiko penularan HIV. Ketujuh atribut tersebut merepresentasikan tiga dimensi utama. Dimensi pertama adalah karakteristik demografis yang terdiri atas *sex*, *age*, dan *educ*. Dimensi kedua adalah perilaku seksual berisiko yang direpresentasikan oleh atribut *consiscondomcs* dan *anystd*. Sementara itu, dimensi ketiga berkaitan dengan persepsi risiko serta paparan terhadap upaya pencegahan HIV, yang direpresentasikan oleh atribut *feelriskcondom* dan *anyintervention*.

Pemilihan ketiga dimensi tersebut didasarkan pada pertimbangan bahwa risiko penularan HIV dipengaruhi oleh berbagai faktor yang saling berkaitan, tidak hanya perilaku seksual, tetapi juga karakteristik demografis serta persepsi dan keterlibatan individu dalam program pencegahan HIV. Pertimbangan tersebut sejalan dengan temuan dalam berbagai penelitian epidemiologi HIV yang menunjukkan bahwa faktor-faktor tersebut berkontribusi terhadap risiko penularan HIV dan dapat dimanfaatkan sebagai atribut prediktor dalam proses klasifikasi.

Tabel 1 Fitur yang Digunakan

No	Fitur	Deskripsi
1	Sex	Jenis kelamin responden (laki-laki/perempuan)
2	Age	Usia responden (numerik)
3	Educ	Tingkat Pendidikan responden
4	Consiscondomcs	Konsistensi penggunaan kondom pada hubungancseksual
5	Anystd	Riwayat pendidikan menular seksual
6	Feelriskcondom	Perasaan berisiko tertular HIV karena tidak selalu menggunakan kondom
7	Anyintervention	Keikut sertaan dalam program/intervensi pencegahan HIV
8	Testresults (target)	Hasil tes HIV responden

Data Cleaning dan Identifikasi Missing Value

Sebelum dilakukan proses imputasi pada data latih, terlebih dahulu dilakukan pemeriksaan terhadap kemungkinan adanya kode yang tidak valid pada setiap atribut yang digunakan. Pemeriksaan ini mencakup pengecekan terhadap kode 8 dan 9 yang pada data survei umumnya digunakan untuk menunjukkan jawaban seperti “tidak berlaku”, “tidak tahu”, atau “tidak menjawab”. Berdasarkan hasil pemeriksaan, tidak ditemukan

adanya kode 8 maupun 9 pada ketujuh atribut yang digunakan dalam penelitian.

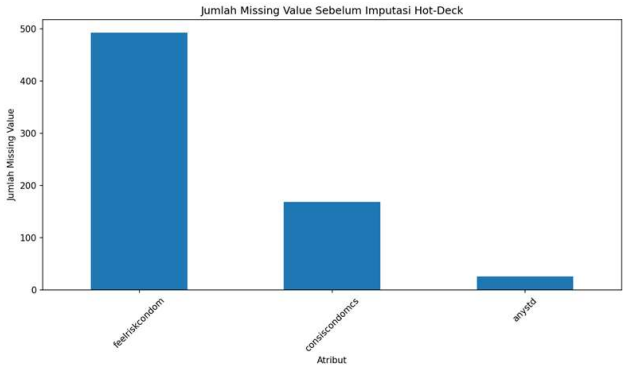
Namun, ditemukan anomali pada atribut *feelriskcondom*, yaitu adanya beberapa nilai yang berada di luar rentang nilai biner yang seharusnya, yaitu 0 dan 1. Nilai-nilai tersebut kemudian dikonversi menjadi *missing value* agar tidak memengaruhi proses pengolahan dan pemodelan data. Penyesuaian ini dilakukan karena nilai di luar rentang kategori yang telah ditentukan dapat dianggap sebagai nilai yang valid apabila tidak dilakukan pemeriksaan terlebih dahulu. Kondisi tersebut berpotensi memengaruhi pola yang terbentuk pada data dan hasil klasifikasi menggunakan algoritma *Random Forest*. Oleh karena itu, nilai yang tidak sesuai dengan ketentuan kategori atribut dikonversi menjadi *missing value* sebelum dilakukan proses imputasi.

Tabel 2. Missing Value pada Data Latih Sebelum Imputasi

No	Fitur	Jumlah Missing Value	Persentase (%)
1	Sex	0	0,00
2	Age	0	0,00
3	Educ	0	0,00
4	Consiscondomcs	132	10,86
5	Anystd	20	1, 64
6	Feelriskcondom	386	31, 74
7	Anyintervention	0	0,00

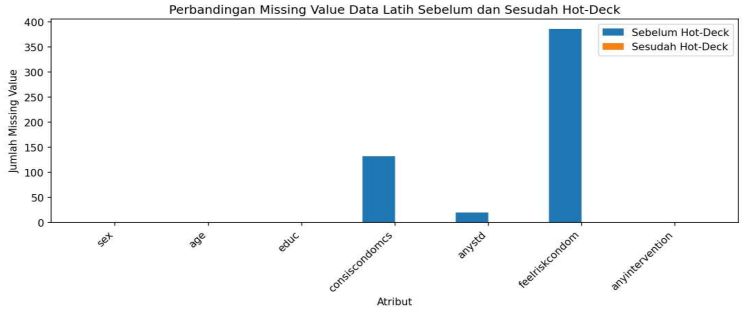
Berdasarkan Tabel di atas, dapat diketahui bahwa tingkat ketidaklengkapan data berbeda pada setiap atribut. Atribut *feelriskcondom* memiliki jumlah *missing value* paling tinggi, yaitu sebanyak 386 dari 1.216 data latih atau sebesar 31,73%. Selanjutnya, atribut *consiscondomcs* memiliki 132 *missing value* atau sebesar 10,86%, sedangkan atribut *anystd* memiliki 20 *missing value* atau sebesar 1,64%. Sementara itu, empat atribut lainnya, yaitu *sex*, *age*, *educ*, dan *anyintervention*, tidak memiliki *missing value* sehingga seluruh datanya tercatat lengkap.

Secara keseluruhan, jumlah *missing value* yang ditemukan pada data latih adalah sebanyak 538 data, sedangkan pada data uji terdapat 150 *missing value*. Nilai yang hilang tersebut selanjutnya menjadi fokus dalam tahap imputasi menggunakan metode *Hot-Deck* untuk melengkapi data sebelum digunakan dalam proses klasifikasi.



Gambar 1. Jumlah Missing Value Sebelum Imputasi Hot-Deck

Tingginya jumlah *missing value* pada atribut *feelriskcondom* diduga berkaitan dengan karakteristik pertanyaan yang bersifat pribadi dan sensitif, sehingga sebagian responden kemungkinan tidak memberikan jawaban. Apabila kondisi tersebut tidak ditangani, keberadaan nilai yang hilang dapat memengaruhi proses pengolahan data dan berpotensi menyebabkan sebagian informasi tidak dapat dimanfaatkan secara optimal dalam proses klasifikasi. Oleh karena itu, dilakukan proses *Hot-Deck Imputation* untuk mengisi nilai yang hilang berdasarkan informasi dari data responden lain yang memiliki karakteristik serupa. Setelah proses imputasi diterapkan, seluruh atribut pada data latih dan data uji telah memiliki nilai yang lengkap atau tidak ditemukan lagi *missing value*.



Gambar 2. Missing Value Sebelum & Sesudah

Hasil Imputasi Hot-Deck

Mekanisme *Hot-Deck Imputation* yang diterapkan dalam penelitian ini dilakukan dengan menggantikan setiap nilai yang hilang menggunakan nilai valid yang dipilih secara acak dari data donor pada atribut yang sama. Pendekatan ini digunakan karena nilai pengganti berasal langsung dari data yang telah tersedia dan teramati, sehingga dapat mempertahankan karakteristik serta distribusi data asli. Berbeda dengan metode imputasi yang menggunakan nilai statistik, seperti rata-rata atau median, metode *Hot-Deck* tidak menghasilkan nilai baru yang bersifat artifisial, melainkan menggunakan nilai aktual

yang terdapat pada data. Dengan demikian, proses imputasi diharapkan dapat melengkapi data yang hilang tanpa mengubah karakteristik data secara signifikan.

Tabel 3. Hasil Imputasi Hot-Deck

Variable	Nilai Sebelum	Nilai Sesudah
Consiscondomcs	NaN	4
Consiscondomcs	NaN	2
Feelriskcondom	NaN	0
Feelriskcondom	NaN	0
Feelriskcondom	NaN	1
Anystd	NaN	0

Berdasarkan Tabel di atas, dapat dilihat bahwa nilai pengganti yang dihasilkan melalui proses *Hot-Deck Imputation* tetap berada dalam rentang kategori yang sesuai dengan atribut masing-masing. Sebagai contoh, atribut *consiscondomcs* tetap memiliki nilai pada rentang kategori 1 hingga 4, sedangkan atribut *feelriskcondom* dan *anystd* tetap berada pada kategori 0 dan 1. Hal ini menunjukkan bahwa metode *Hot-Deck Imputation* mampu mempertahankan struktur dan karakteristik kategori pada setiap atribut yang digunakan. Dengan demikian, proses imputasi dapat melengkapi nilai yang hilang tanpa mengubah makna kategori pada data, sehingga integritas dan konsistensi data tetap terjaga sebelum digunakan pada tahap pemodelan selanjutnya.

Transformasi Lanjutan

Berdasarkan hasil pemeriksaan terhadap tipe data dan rentang nilai pada ketujuh atribut yang digunakan, diketahui bahwa seluruh atribut telah memiliki format data numerik atau ordinal sejak awal. Atribut *sex* memiliki rentang nilai 1–2, sedangkan atribut *age* memiliki rentang usia 15–65 tahun. Atribut *educ* dan *consiscondomcs* merupakan data berskala ordinal dengan rentang nilai 1–4. Sementara itu, atribut *anystd*, *feelriskcondom*, dan *anyintervention* merupakan variabel biner yang memiliki nilai 0 dan 1.

Hasil pemeriksaan juga menunjukkan bahwa tidak terdapat atribut dalam bentuk teks atau data kategorikal nominal yang memerlukan proses *encoding*. Selain itu, proses normalisasi maupun standardisasi tidak diperlukan sebelum data digunakan dalam pemodelan menggunakan algoritma *Random Forest*. Hal ini dikarenakan *Random Forest* merupakan algoritma berbasis pohon keputusan yang tidak sensitif terhadap perbedaan skala antarvariabel. Dengan demikian, data yang telah melalui proses pemeriksaan dan imputasi dapat langsung digunakan pada tahap pemodelan klasifikasi.

B. Hasil Klasifikasi Random Forest

Pembentukan model klasifikasi dalam penelitian ini menggunakan algoritam Random Forest dengan menerapkan beberapa konfigurasi parameter. Nilai parameter *n_estimators* ditentukan sebesar 100 yang berarti model terdiri atas 100 pohon keputusan, sedangkan parameter *max_depth* menggunakan nilai *None* sehingga kedalaman pohon tidak dibatasi. Kriteria yang digunakan dalam proses pemisahan *node* adalah *Gini Index*. Parameter *class_weight* diatur menggunakan *balanced* untuk membantu menyesuaikan bobot pada masing-masing kelas berdasarkan proporsi data yang tersedia. Sementara itu, parameter *random_state* ditetapkan sebesar 42 untuk memastikan hasil eksperimen dapat direproduksi.

Tahap pelatihan model menggunakan sebanyak 1.216 data latih yang sebelumnya telah melalui proses imputasi dengan metode Hot-Deck. Selanjutnya, model diuji menggunakan 305 data uji yang telah melalui proses imputasi dengan memanfaatkan donor yang berasal dari data latih. Dengan konfigurasi tersebut, model *Random Forest* digunakan untuk melakukan proses klasifikasi berdasarkan atribut yang telah melalui tahap prapemrosesan dan imputasi.

Confusion Matrix

Tabel 4. Confusion Matrix Hasil Klasifikasi

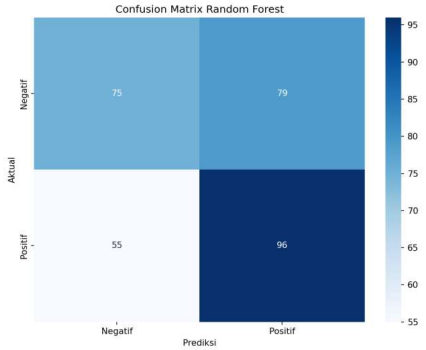
	Prediksi Kelas 0	**Prediksi Kelas 1**
Aktual Kelas 0 (TN + FP)	75 (TN)	79 (FP)
Aktual Kelas 1 (FN + TP)	55(FN)	96 (TP)

Berdasarkan Tabel di atas, hasil *confusion matrix* menunjukkan bahwa dari 154 responden yang secara aktual termasuk dalam kelas negatif (kelas 0), sebanyak 75 responden berhasil diklasifikasikan dengan tepat sebagai negatif (*True Negative*), sedangkan 79 responden lainnya diklasifikasikan secara keliru sebagai positif (*False Positive*). Sementara itu, dari 151 responden yang secara aktual termasuk dalam kelas positif (kelas 1), model berhasil mengklasifikasikan 96 responden dengan tepat sebagai positif (*True Positive*). Namun, sebanyak 55 responden lainnya tidak berhasil terdeteksi oleh model sebagai positif dan justru diklasifikasikan sebagai negatif (*False Negative*).

Berdasarkan hasil yang diperoleh, model menunjukkan kemampuan yang lebih baik dalam mengidentifikasi data pada kelas positif dibandingkan dengan data kelas negatif. Hal ini dapat dilihat dari jumlah *True Positive* yang lebih tinggi dibandingkan dengan *True Negative*. Namun, masih terdapat sejumlah kesalahan klasifikasi yang ditunjukkan

oleh nilai *False Positive* dan *False Negative*, sehingga diperlukan evaluasi lebih lanjut menggunakan metrik kinerja lainnya untuk mengetahui kemampuan model secara keseluruhan.

Perbandingan jumlah *True Positive* dan *True Negative* menunjukkan bahwa model memiliki kecenderungan yang relatif lebih baik dalam mengenali responden yang lebih banyak diklasifikasikan ke dalam kelas positif dibandingkan dengan kelas negatif. Namun, jumlah False Positif yang relative tinggi menunjukkan bahwa masih terdapat sejumlah data dari kelas negative yang keliru diklasifikasikan sebagai kelas positif yaitu sebanyak 79 responden, menunjukkan bahwa model masih cukup sering mengklasifikasikan responden yang sebenarnya termasuk dalam kelas negatif sebagai kelas positif. Kondisi ini menunjukkan bahwa kemampuan model dalam membedakan kelas negatif dan positif masih perlu diperhatikan, khususnya dalam mengurangi kesalahan prediksi pada kelas negatif.



Gambar 3. Confusion Matrix

Metrik Evaluasi

Kinerja model dievaluasi secara kuantitatif dengan menggunakan empat metrik utama, yaitu Accuracy, Precision, Recall, dan F1-Score. Setiap metrik digunakan untuk mengukur aspek yang berbeda dalam menilai kemampuan model dalam menghasilkan prediksi yang tepat.

Tabel 5. Evaluasi Model Random Forest dengan Imputasi Hot-deck

No	Metrik	Nilai	Persentase (%)
1.	Accuracy	0,5607	56,07%
2.	Precision	0,5486	54,86%
3.	Recall	0,6358	63,58%
4.	F1-Score	0,5890	58,90%

Berdasarkan Tabel di atas, model *Random Forest* dengan penerapan imputasi *Hot-*

Deck menghasilkan nilai *Accuracy* sebesar 56,07%, yang menunjukkan proporsi keseluruhan data uji yang berhasil diklasifikasikan dengan benar oleh model. Sementara itu, nilai *Precision* sebesar 54,86% menunjukkan bahwa dari seluruh responden yang diprediksi termasuk ke dalam kelas positif, sekitar 54,86% di antaranya sesuai dengan kelas aktualnya. Nilai *Recall* sebesar 63,58% menunjukkan kemampuan model dalam mengidentifikasi responden yang secara aktual dikategorikan sebagai kelas positif. Nilai *Recall* yang lebih tinggi dibandingkan *Precision* menunjukkan bahwa model relatif lebih mampu mendeteksi kelas positif, meskipun masih terdapat sejumlah data positif yang tidak berhasil teridentifikasi. Dalam konteks klasifikasi terkait kesehatan, kemampuan mendeteksi kelas positif menjadi salah satu aspek yang perlu diperhatikan karena kesalahan klasifikasi berupa *False Negative* dapat menyebabkan individu yang sebenarnya termasuk dalam kelas positif dikategorikan sebagai negatif.

Nilai *F1-Score* sebesar 58,90% menunjukkan keseimbangan antara *Precision* dan *Recall* dalam kinerja model. Nilai tersebut mengindikasikan bahwa model belum mencapai performa yang optimal dalam menyeimbangkan kemampuan untuk menghasilkan prediksi positif yang tepat dengan kemampuan dalam mendeteksi seluruh data yang sebenarnya termasuk dalam kelas positif. Jika dibandingkan dengan *baseline* sederhana berupa prediksi berdasarkan kelas mayoritas, yang secara teoritis menghasilkan akurasi sebesar 50,49%, model *Random Forest* yang dibangun masih menunjukkan kinerja yang lebih baik dengan nilai *Accuracy* sebesar 56,07%. Meskipun peningkatannya relatif terbatas, hasil tersebut menunjukkan bahwa model mampu memberikan prediksi yang lebih baik dibandingkan pendekatan yang hanya memprediksi berdasarkan kelas mayoritas.

Classification Report per Kelas

Selain melakukan evaluasi kinerja model secara keseluruhan, dilakukan pula evaluasi berdasarkan masing-masing kelas untuk mengetahui kemampuan model dalam mengklasifikasikan setiap kategori. Evaluasi ini dilakukan menggunakan *classification report* yang memuat nilai *Precision*, *Recall*, dan *F1-Score* pada setiap kelas.

Tabel 6. Classification Report per Kelas

Kelas	Precision	Recall	F1-Score	Support
0 (Negatif)	0,58	0,49	0,53	154
1 (Positif)	0,55	0,64	0,59	151
Macro Avg	0,56	0,56	0,56	305

Weighted Avg	0,56	0,56	0,56	305

Berdasarkan Tabel di atas, model menunjukkan bahwa kelas 1 (positif) memiliki nilai *F1-Score* yang sedikit lebih tinggi dibandingkan dengan kelas 0 (negatif), yaitu masing-masing sebesar 0,59 dan 0,53. Perbedaan nilai tersebut menunjukkan bahwa kemampuan model dalam mengenali kedua kelas relatif tidak jauh berbeda. Hal ini juga sejalan dengan komposisi data uji yang relatif seimbang, yaitu terdiri dari 154 responden pada kelas negatif dan 151 responden pada kelas positif. Dengan demikian, tidak terlihat adanya ketimpangan performa yang terlalu besar dalam klasifikasi kedua kelas, meskipun model menunjukkan kinerja yang sedikit lebih baik dalam mengenali kelas positif.

C. Pembahasan

Berdasarkan hasil pengujian menggunakan algoritma *Random Forest*, diperoleh nilai *Accuracy* sebesar 56,07%, *Precision* sebesar 54,86%, *Recall* sebesar 63,58%, dan *F1-Score* sebesar 58,90%. Selisih antara nilai akurasi model dan *baseline* berupa prediksi berdasarkan kelas mayoritas, yaitu sebesar 50,49%, menunjukkan bahwa peningkatan kinerja model masih relatif terbatas. Hasil tersebut mengindikasikan bahwa ketujuh atribut demografis dan perilaku yang digunakan dalam penelitian ini belum sepenuhnya mampu memberikan prediksi dengan tingkat ketepatan yang tinggi terhadap status hasil pemeriksaan HIV. Oleh karena itu, terdapat kemungkinan bahwa penggunaan atribut tambahan yang lebih relevan atau informatif dapat dipertimbangkan untuk meningkatkan kemampuan model dalam melakukan klasifikasi.

Nilai *Recall* yang lebih tinggi dibandingkan dengan *Precision* menunjukkan bahwa model relatif lebih baik dalam mengidentifikasi responden yang termasuk ke dalam kelas positif dibandingkan dalam menghindari kesalahan klasifikasi pada kelas negatif. Hasil tersebut sejalan dengan *Confusion Matrix*, yang menunjukkan bahwa jumlah *False Positive* sebanyak 79 responden lebih tinggi dibandingkan dengan jumlah *False Negative* sebanyak 55 responden. Kondisi ini menunjukkan bahwa model memiliki kecenderungan untuk mengklasifikasikan lebih banyak responden ke dalam kelas positif, meskipun sebagian di antaranya secara aktual termasuk dalam kelas negatif. Dengan demikian, model cenderung lebih sensitif dalam mengidentifikasi kelas positif, meskipun masih terdapat kesalahan klasifikasi berupa False Positif yang cukup tinggi.

Keterbatasan performa model yang diperoleh dalam penelitian ini kemungkinan dipengaruhi oleh karakteristik atribut yang digunakan. Atribut *sex*, *age*, *educ*,

consiscondomcs, *anystd*, *feelriskcondom*, dan *anyintervention* merupakan variabel yang menggambarkan karakteristik demografis serta perilaku responden, sehingga belum secara langsung merepresentasikan kondisi klinis atau status infeksi HIV. Atribut-atribut tersebut juga tidak mencakup indikator medis secara langsung, seperti hasil pemeriksaan laboratorium atau informasi biomedis lainnya yang dapat memberikan gambaran lebih spesifik mengenai status infeksi HIV.

Selain itu, tingginya jumlah *missing value* pada atribut *feelriskcondom* sebelum dilakukan proses imputasi juga menunjukkan adanya keterbatasan informasi pada data yang digunakan. Meskipun metode *Hot-Deck Imputation* telah diterapkan untuk mengisi nilai yang hilang dengan mempertahankan karakteristik dan distribusi data yang tersedia, proses imputasi tetap tidak dapat sepenuhnya menggantikan informasi asli yang tidak diperoleh dari responden. Kondisi tersebut dapat menjadi salah satu faktor yang berpengaruh terhadap tingkat ketepatan model dalam melakukan klasifikasi.

Meskipun performa klasifikasi yang diperoleh masih tergolong moderat, penelitian ini menunjukkan bahwa metode *Hot-Deck Imputation* dapat diterapkan untuk menangani *missing value* pada dataset penyakit menular. Proses imputasi berhasil melengkapi seluruh nilai yang hilang tanpa menghapus baris data, sehingga seluruh 1.521 responden tetap dapat dimanfaatkan dalam proses klasifikasi. Hasil tersebut menunjukkan bahwa metode *Hot-Deck Imputation* dapat menjadi salah satu alternatif teknik *pre-processing* dalam menangani permasalahan *missing value*, khususnya pada data kesehatan yang memiliki kemungkinan mengalami ketidaklengkapan data. Selanjutnya, penelitian dapat dikembangkan dengan menambahkan atribut yang lebih relevan, menggunakan sumber data yang lebih lengkap, atau membandingkan metode imputasi dan algoritma klasifikasi lainnya untuk memperoleh hasil prediksi yang lebih optimal.

5. KESIMPULAN DAN SARAN

Metode Hot-Deck Imputation berhasil diterapkan untuk menangani seluruh *missing value* pada dataset penyakit menular (HIV), yaitu sebanyak 538 nilai pada data latih dan 150 nilai pada data uji (total 688 nilai), tanpa mengurangi jumlah baris data yang digunakan dalam proses analisis, sehingga seluruh 1.521 responden tetap dapat dimanfaatkan dalam proses klasifikasi. Model klasifikasi Random Forest yang dibangun menggunakan dataset hasil imputasi Hot-Deck model yang digunakan memperoleh nilai Accuracy sebesar 56,07%, Precision sebesar 54,86%, Recall sebesar 63,58%, dan F1-

Score sebesar 58,90%. Nilai Recall yang lebih tinggi dibandingkan Precision menunjukkan bahwa model relatif lebih baik dalam mengidentifikasi responden yang termasuk ke dalam kelas positif (berisiko) dibandingkan dalam menghindari kesalahan klasifikasi pada kelas negatif. Performa klasifikasi yang diperoleh masih tergolong moderat, dengan peningkatan yang relatif terbatas dibandingkan baseline prediksi kelas mayoritas sebesar 50,49%. Hal ini mengindikasikan bahwa atribut demografis dan perilaku yang digunakan belum sepenuhnya mampu memberikan prediksi dengan tingkat ketepatan yang tinggi terhadap status hasil pemeriksaan HIV. Meskipun demikian, penelitian ini membuktikan bahwa metode Hot-Deck Imputation dapat diterapkan sebagai tahap pre-processing untuk menangani missing value pada dataset penyakit menular tanpa perlu menghapus data, sehingga dapat menjadi salah satu alternatif teknik penanganan missing value pada data kesehatan yang memiliki kemungkinan mengalami ketidaklengkapan data.

DAFTAR REFERENSI

Ardiansyah, D. N. (2025). *Pemodelan Data Perilaku Seksual dan Penyalahgunaan Narkoba dengan Machine Learning*. Universitas Jenderal Achmad Yani.

Baratawidjaja. (2024). Skrining Penyakit Menular di RSUD RA Kartini Jepara. *Hikmah Journal of Health*, *2*(1).

Bertha, S. (2019). Konfirmasi Lima Faktor yang Berpengaruh terhadap Perilaku Remaja dalam Pencegahan HIV/AIDS. *Jurnal Ilmiah Kebidanan Indonesia*, *9*(4).

Centers for Disease Control and Prevention. (2026). *Memahami Pengalaman Pengungkapan Status dan Penginformasian Medikasi HIV/AIDS Oleh Pengasuh Kepada Anak dengan HIV/AIDS*. Retrieved from https://ejournal3.undip.ac.id/index.php/interaksi-online/article/download/41321/29922

Dedot. (2023). Penerapan Algorityma Decision Tree Berbasis Pohon Keputusan Dalam Klasifikasi Penyakit Jantung. *Jurnal Mahasiswa Teknik Informatika*, *7*(6).

Delen, D. (2026). Tahap-tahap Data Mining. Retrieved January 21, 2026, from https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/

Efron, B., Halloran, E., & Holmes, S. (2026). The Following Article, Which Appreared in Number 14. Retrieved January 21, 126 C.E., from https://www.pnas.org

Fadillah, I. J., & S.Muchlisoh. (2002). *Perbandingan Metode Hot-Deck Imputation dan Metode KNNI Dalam Mengatasi Missing Values*. *2019*(1). https://doi.org/10.34123

Han, J., Kamber, M., & Pei, J. (2016). *Data Mining: Concepts and Techniques*.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Hoboken, NJ, USA: Wiley.

Rhaudatunnisa, T., & Wilantika, N. (2019). Performance Comparison of Hot-Deck Imputation, K-Nearest Neighbor Imputation, and Predictive Mean Matching in Missing Value Handling. *SUSENAS Kor Dataset*.

Tan, P. N., Steinbach, M., & V.Kumar. (2016). Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Negihbor. *Jurnal Teknik Pomits*, *2*(1).

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Boca Raton, FL, USA: CRC Press.